CUNY SPS DATA 621 - CTG5 - HW3

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

April 10th, 2019

Contents

1 DATA EXPLORATION

Relocating to a new city or state can be very stressful. In addition to the stress of packing and moving, you may also be nervous about moving to an unfamiliar area. To better understand their new community, some new residents or people interested in moving to a new city choose to review crime statistics in and around their neighborhood. Crime rate may also influence where people choose to live, raise their families and run their businesses; many potential new residents steer clear of cities with higher than average crime rates.

Data was collected in order to predict whether the nighborhood will be at risk for high crime levels. For each neghborhood the response variable, `target`, represents whetever the crime rate is above the crime rate or not. In addition to that 13 predictor variables were collected representing each neighborhood's: large lots, non-retail business acres, nitrogen oxides concentration, average number of rooms per dwelling, proportion of owner-occupied units, distances to five Boston employment centers, accessibility to radial highways, property tax rate, pupil-teacher ration. The evaluation data contains the same 13 predictor variables and no target variable so it will be impossible to check the accuracy of our predictions from the testing data.

VARIABLE NAME	DEFINITION	TYPE
target	whether the crime rate is above the median crime rate (1) or not (0)	response variable
zn	proportion of residential land zoned for large lots (over 25000 square feet)	predictor variable
indus	proportion of non-retail business acres per suburb	predictor variable
chas	a dummy var. for whether the suburb borders the Charles River (1) or not (0)	predictor variable
nox	nitrogen oxides concentration (parts per 10 million)	predictor variable
rm	average number of rooms per dwelling	predictor variable
age	proportion of owner-occupied units built prior to 1940	predictor variable
dis	weighted mean of distances to five Boston employment centers	predictor variable
rad	index of accessibility to radial highways	predictor variable
tax	full-value property-tax rate per $10,000	predictor variable
ptratio	pupil-teacher ratio by town	predictor variable
black	$1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town	predictor variable
lstat	lower status of the population (percent)	predictor variable
medv	median value of owner-occupied homes in $1000s	predictor variable

1.1 Summary Statistics

Looking at the Table 2, we can see that `chas` and `target` are binary variable. 49% of all target varibles are 0s. There are outliers present in `zn`, `lstat`, `medv` and `dis`.

1.2 Shape of Predictor Distributions

Figure. 1 shows that the distribution of most of the variables seems skewed. There are some outliers in the right tail of `tax` , `rad`, `medv`, `lstat`, `dis` and left tail of `ptratio`.

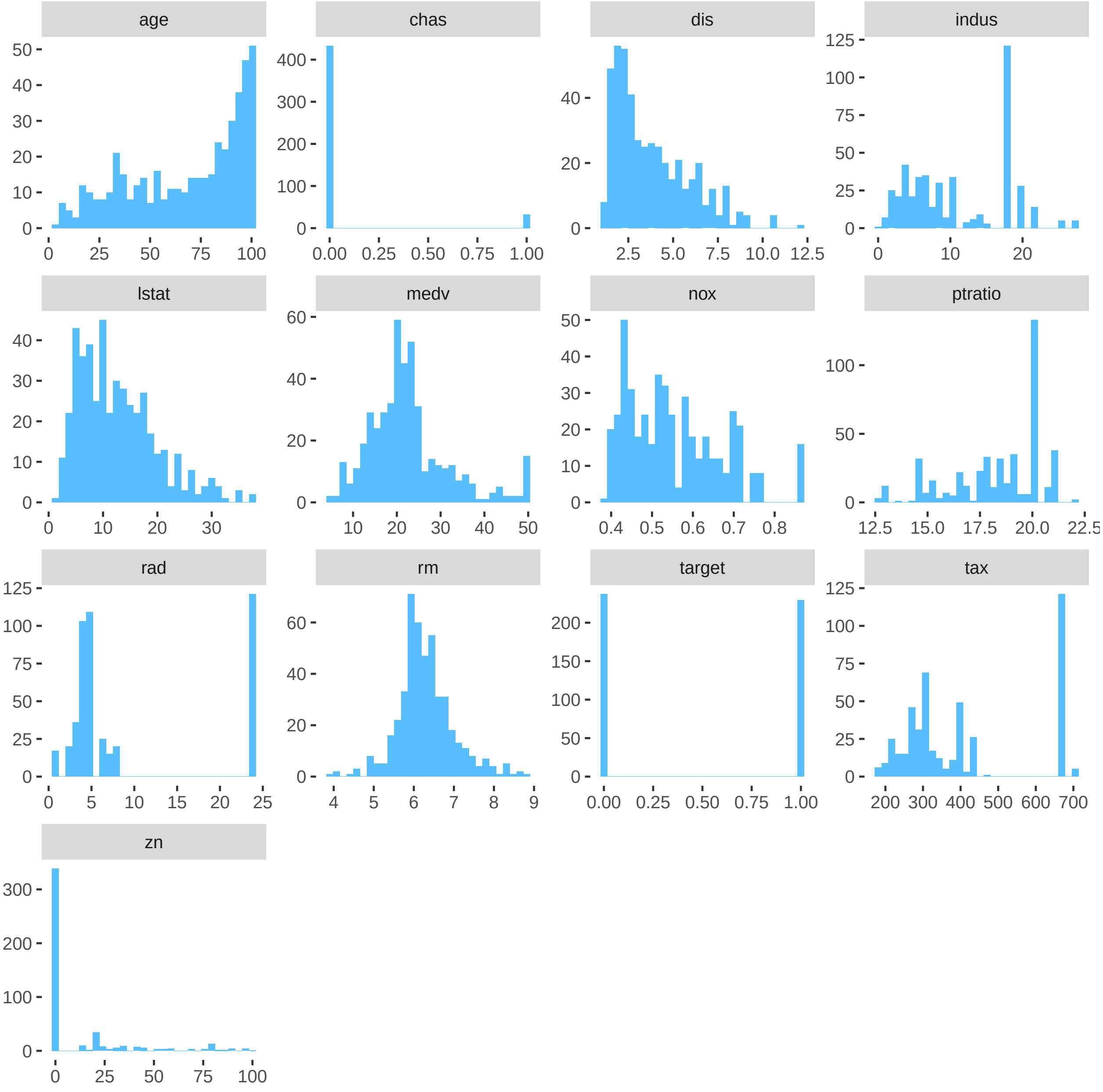

Figure 1: Data Distributions

Table 2: Summary statistics

	n	min	mean	median	max	sd
zn	466	0.0000	11.5772532	0.00000	100.0000	23.3646511
indus	466	0.4600	11.1050215	9.69000	27.7400	6.8458549
chas	466	0.0000	0.0708155	0.00000	1.0000	0.2567920
nox	466	0.3890	0.5543105	0.53800	0.8710	0.1166667
rm	466	3.8630	6.2906738	6.21000	8.7800	0.7048513
age	466	2.9000	68.3675966	77.15000	100.0000	28.3213784
dis	466	1.1296	3.7956929	3.19095	12.1265	2.1069496
rad	466	1.0000	9.5300429	5.00000	24.0000	8.6859272
tax	466	187.0000	409.5021459	334.50000	711.0000	167.9000887
ptratio	466	12.6000	18.3984979	18.90000	22.0000	2.1968447
lstat	466	1.7300	12.6314592	11.35000	37.9700	7.1018907
medv	466	5.0000	22.5892704	21.20000	50.0000	9.2396814
target	466	0.0000	0.4914163	0.00000	1.0000	0.5004636

1.3 Outliers

Figure. 2 shows that there are also a large number of outliers that need to be accounted for, most prevalently in `zn` and `medv` based off of the boxplots below. Since `tax` variable has values which are very large compared to other variables in the dataset, it was scaled to fit the boxplot.

1.4 Missing Values

Figure. 3 displays of all the observations gathered across these thirteen variables, there are 0 missing values.

1.5 Linearity

Each variable was plotted against the target variable in order to determine at a glance which had the most potential linearity before the dataset was modified.

As can be observed in Figure. 4, the most influential variables are the ones previously discussed to have severe outliers and skew, and their linear relationship is negative - the higher the variable, the lower the target wins.

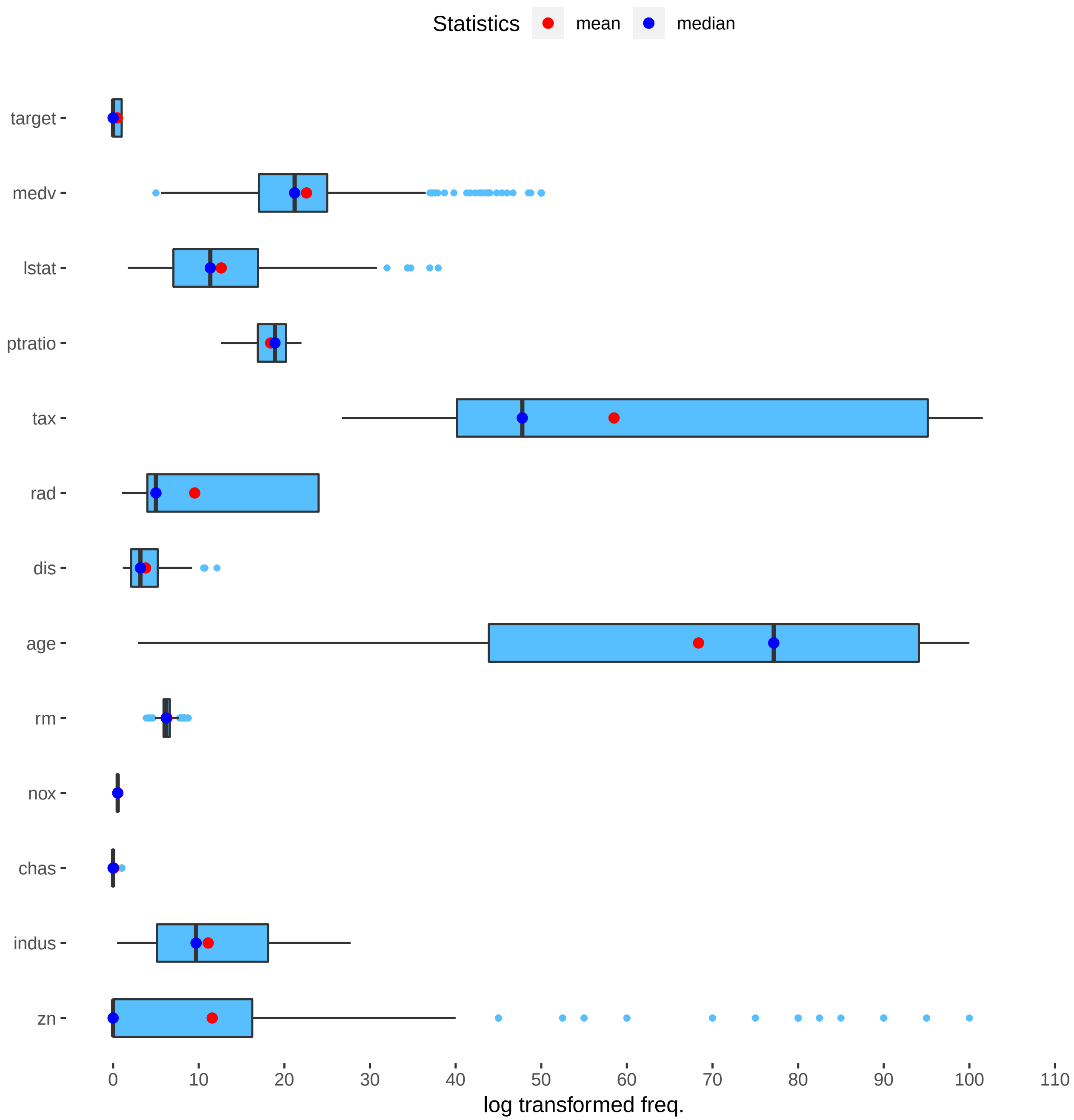

Figure 2: Boxplots highlighting many outliers in the data.

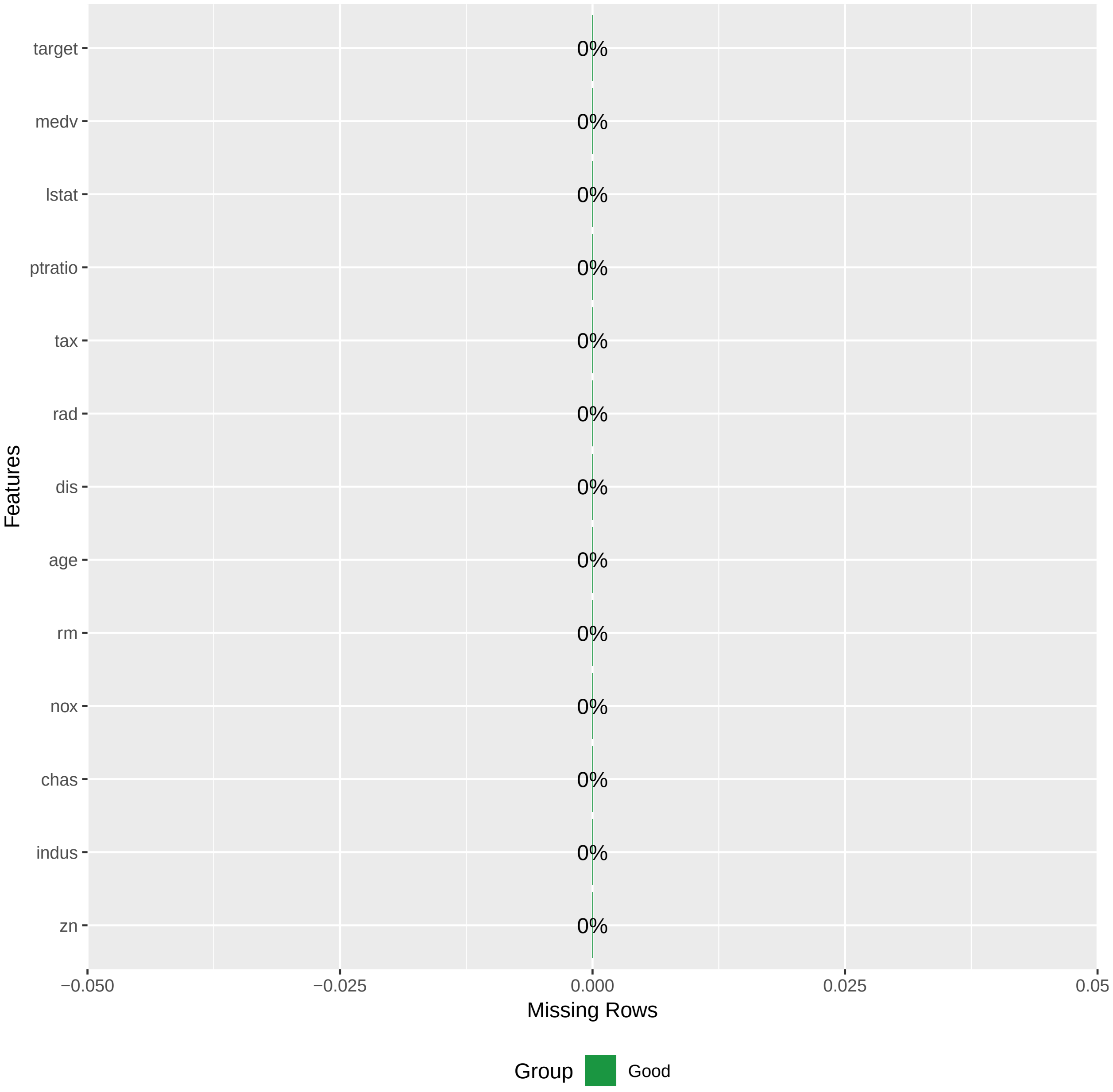

Figure 3: Missing values

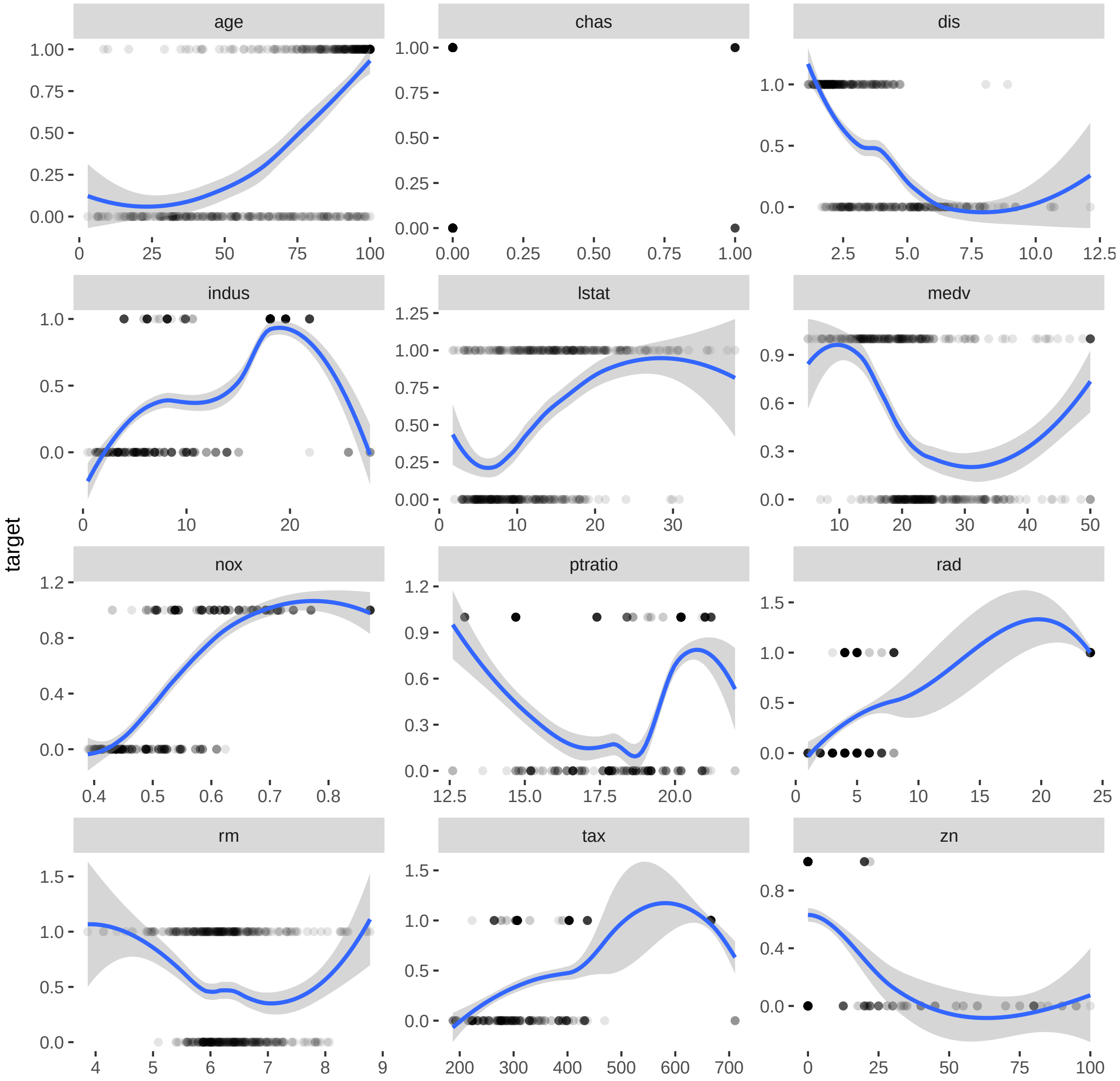

Figure 4: Linear relationships between each predictors and the target